Australian Bureau of Statistics

Research Paper

Synthetic Microdata - A Possible Dissemination Tool

Australia

2019

1351.0.55.163

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30AM (CANBERRA TIME) FRI 23 AUG 2019

ABS Catalogue No. 1351.0.55.163

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

Reissue

Research Paper

Synthetic Microdata – A Possible Dissemination Tool

Methodology Transformation Branch

Methodology Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 23 AUG 2019

SYNTHETIC BUSINESS MICRODATA : A POSSIBLE DISSEMINATION TOOL

Chien-Hung Chien[1,2], A.H. Welsh[3], and John D Moore[1]

[1]Australian Bureau of Statistics (ABS)

[2]ANU Mathematical Science Institute

[3]ANU Research School of Finance, Actuarial Studies & Statistics

ABSTRACT

Enhancing microdata access is one of the strategic priorities for the Australian Bureau of Statistics (ABS) in its transformation program. However, balancing the trade-off between enhancing data access and protecting confidentiality is a delicate act. The ABS could use synthetic data to make its business microdata more accessible for researchers to inform decision making while maintaining confidentiality. This study explores the synthetic data approach for the release and analysis of business data. Australian businesses in some industries are characterised by oligopoly or duopoly. This means the existing microdata protection techniques such as information reduction or perturbation may not be as effective as for household microdata. The research focuses on addressing the following questions: Can a synthetic data approach enhance microdata access for the longitudinal business data? What is the utility and protection trade-off using the synthetic data approach? The study compares confidentialised input and output approaches for protecting confidentiality and analysing Australian microdata from business survey or administrative data sources.

CONTENTS

1 INTRODUCTION

Statistical agencies are constantly facing decisions on how to best balance the trade-off between protecting data confidentiality and providing greater access to the valuable data they collect to inform decision making. Clause 7 of the Statistics Determination 1983 ensures safe access to ABS data in the form of unidentified individual statistical records (microdata), for research and analysis purposes. Clause 7 stipulates that "the information is disclosed in a manner that is not likely to enable the identification of the particular person or organization to which it relates" The Australian Government (1983). Protections are important for producing high quality statistics. However, protections have to be balanced with appropriate levels of data access and dissemination. As economist George Stigler pointed out in 1980, data is both a private and public good. On the one hand, statistical agencies must protect confidentiality, but at the same time they also need to ensure that data is accessible so that it can be used to inform decisions that have significant impact on the public interest (Abowd and Schmutte, 2015).

The ABS has increasingly emphasised providing better access to microdata for research. The ABS uses the Five Safes Framework to ensure microdata can be used appropriately by taking into consideration safe people, projects, settings, data and output (ABS, 2016, Desai et al., 2016). The ABS provides three types of microdata products - TableBuilder, Confidentialised Unit Record Files or CURFs and detailed microdata (ABS, 2017). The microdata access methods include both remote or on-site depending on the type of microdata product. Users can analyse highly detailed microdata in the on-site ABS Data Laboratory or DataLab environment. For access from user's own environment, the ABS has provided a suite of microdata products such as TableBuilder (tabulation of Census or surveys) and Remote Access DataLab (analysing the more detailed CURFs). For microdata access, researchers can download basic CURFs for analysis in their own environment. However, these basic CURFs contain little detail and are reported at a more aggregate level (Tam et al., 2009). These microdata products facilitate research that maximises the value of data for informing decisions of importance to Australia. Examples include Healy et al. (2015), Breunig and Bakhtiari (2013), Blackmore and Nesbitt (2013).

The ABS releases CURFs, using suppression, aggregation, and top and bottom coding methodologies, to enable analysis of microdata (O'Keefe and Shlomo, 2012). However, these techniques can make microdata from business surveys or administrative data sources (or business microdata) less useful because some Australian industries are characterised by oligopoly or duopoly. Useful information is often suppressed or aggregated to avoid re-identification of large businesses. The ABS could consider releasing synthetic datasets for researchers to enhance access to business microdata. Synthetic datasets preserve the relationships between variables so that researchers can make valid inferences about the

target population without accessing the underlying microdata (Loong, 2012). The US Census Bureau uses synthetic data to make its business microdata more accessible to researchers and provides a validation service.

This research explores the use of a synthetic data approach as a possible dissemination tool for Australian business microdata. The first section describes the statistical models and processes to impute missing data. Using imputed data to create synthetic microdata provides the advantages of enhancing utility and protection of the synthetic microdata. The second section describes the different disclosure control approaches including confidentialise input and confidentialise output. The third section provides utility and risk results for the different approaches. The final section contains conclusions.

2 STATISTICAL MODEL AND DATA ANALYSIS

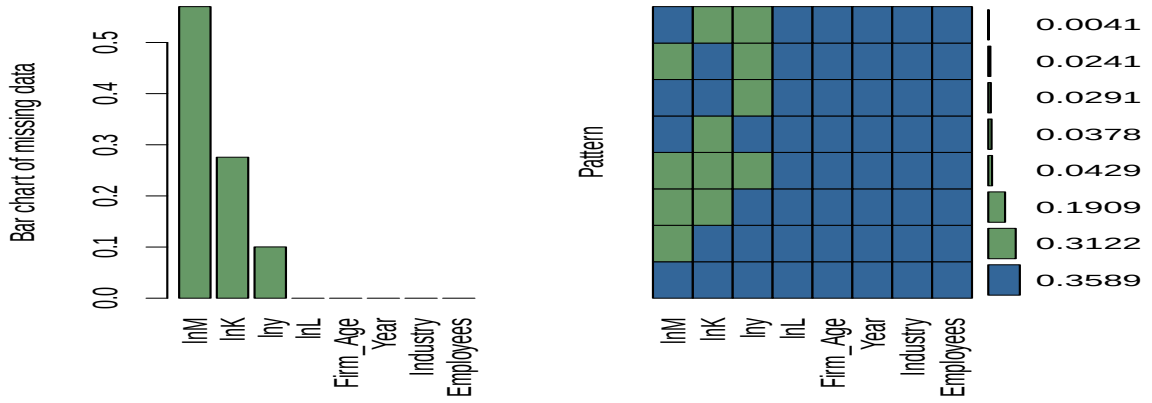
We are interested in preserving the statistical relationships between the variables in the firm production function. The statistical model is specified as:

$$\ln y_{jkt} = \alpha_1 \ln L_{jkt} + \alpha_2 \ln K_{jkt} + \alpha_3 \ln M_{jkt} + \alpha_3 \ln Firm\_Age_{jkt} + \tau_{kt} + \epsilon_{jkt}, \quad (1)$$

where $\ln y_{jkt}$ is the logarithm of total sales adjusted for the repurchase of stocks divided by the total number of employees for firm j in industry k at time t. The logarithm of estimated firm average labour components $\ln L_{jkt}$ for firm j in industry k at time t is derived using the method proposed by Abowd et al. (2002). Details can be found in **?**. The logarithm of capital cost $\ln K_{jkt}$ is the logarithm of the sum of equipment depreciation, business rental expenses and capital investment deductions divided by the total number of employees for firm j in industry k at time t. The logarithm of material costs $\ln M_{jkt}$ is the logarithm of the inputs used in the production process divided by the total number of employees for firm j in industry k and time t. The logarithm of firm age is $\ln Firm\_Age_{jkt}$ for firm j in industry k at time t. We also include time fixed effects τ_{kt} for industry k at time t (Breunig and Wong, 2008, Nguyen and Hansell, 2014, Mare et al., 2016). This gives 15 unknown regression parameters in (1). This study used a one percent stratified sample of business microdata from an expanded prototype dataset ($N > 45000$ firms). Chien and Mayer (2015), Chien et al. (2019) provide more details of the prototype dataset. We simplify notation in (1) by **removing the subscripts**. We also use different fonts i.e., $\boldsymbol{\mathcal{X}}$, to represent observed $N \times 15$ matrix containing all the independent variables in (1). Similarly, we use $\boldsymbol{y}$ to represent the observed vector containing dependent variable in (1).

The prototype sample contains missing values, particularly for material inputs. Figure 1 shows the missing data pattern; the three variables with missing values include ($\ln M$, $\ln K$ and $\ln y$) in descending order.

Figure 1: Missing data pattern



Note. The green tile indicates missing data. The blue tile indicates non missing data. Consider ABS and Patents subfigure at the top left, the left panel is a bar chart showing the propostion of missing data for each variable. The right panel shows the 8 missing data patterns in the data and the proportion of each pattern.

The missing values in the 1% sample are imputed assuming the data are missing at random (MAR). The consequence of this assumption is that missing values can be imputed using models fitted to the observed data (Little and Rubin, 2014). We adapt a similar notation to Reiter (2005a). The experimental dataset consists of $[\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{y}$ is $N \times 1$ vector which includes the dependent variable, and $\boldsymbol{\mathcal{X}}$ is $N \times 15$ matrix which includes all the independent variables in (1). We have imputed the missing variables $\ln y$, $\ln K$ and $\ln M$. We use two Bayesian imputation approaches - Predictive Mean Matching and Expectation Maximisation and Bootstrap to impute the missing data.

The observed dataset consists of two $N \times 16$ matrices, $\boldsymbol{\mathcal{D}} = [\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{\mathcal{X}}$ includes all the independent variables in (1), and the response indicator matrix $\boldsymbol{\mathcal{R}}$ which we use to partition $\boldsymbol{\mathcal{D}}$ into the observed $\boldsymbol{\mathcal{D}}_{obs}$ and the missing $\boldsymbol{\mathcal{D}}_{mis}$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the matrix for imputing missing data in $\ln y$, $\ln K$ and $\ln M$, respectively. So if the missing data variable is $\ln y$ then $\boldsymbol{\mathcal{X}}$ includes all the independent variables in (1). In comparison, if the missing data variable is $\ln K$ then $\boldsymbol{\mathcal{X}}^{(K)}$ includes all the independent variables and $\ln y$ but excludes $\ln K$. If the missing data variable is $\ln M$ then $\boldsymbol{\mathcal{X}}^{(M)}$ includes all the independent variables and $\ln y$ but excludes $\ln M$. We impute the missing values in $\ln y$, $\ln K$ and $\ln M$ separately, using two Bayesian imputation approaches - Predictive Mean Matching (PMM) and Expectation Maximisation and Bootstrap (EMB).

PMM selects from a set of possible donors from the complete cases whose predictive means are closest to that of the missing case (Little, 1988). The value of the selected $\boldsymbol{y}_{obs}$ are then imputed for $\boldsymbol{y}_{mis}$. This method is similar to a hot-deck imputation because it randomly

choose one $\boldsymbol{y}_{imp}$ from nearest neighbour complete cases. Vink et al. (2014) shows that similar to (4), the PMM formula for imputing a target variable $\boldsymbol{y}$ can be expressed as:

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

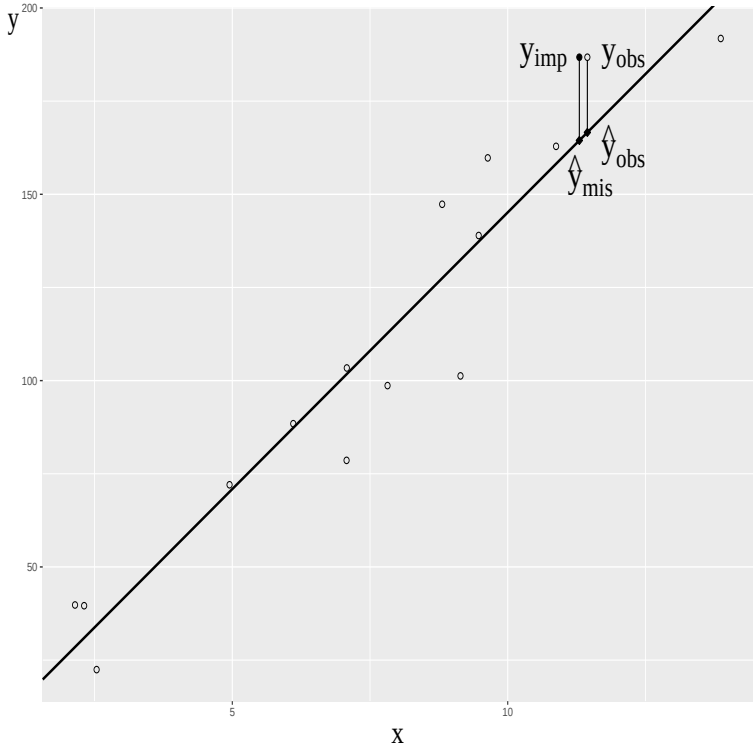
The box 1 describes the concept of the algorithm.

Algorithm 1 PMM algorithm

1: **procedure**
2: use $\boldsymbol{\mathcal{D}}_{obs}$ to estimate $\widehat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$.
3: use draw variance $\tilde{\sigma}^2$ from $\hat{\boldsymbol{\epsilon}}^\top\hat{\boldsymbol{\epsilon}}/A$ where A is χ^2 with $N-k$ with k is the number of parameters.
4: draw $\widetilde{\boldsymbol{\beta}}$ from a multivariate normal distribution centered at $\widehat{\boldsymbol{\beta}}$ with covariance matrix $\tilde{\sigma}^2(\boldsymbol{\mathcal{X}}_{obs}^\top\boldsymbol{\mathcal{X}}_{obs})^{-1}$.
5: calculate $\hat{\boldsymbol{y}}_{obs} = \boldsymbol{\mathcal{X}}_{obs}\widehat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}_{mis} = \boldsymbol{\mathcal{X}}_{mis}\widetilde{\boldsymbol{\beta}}$.
6: **for each** each $\boldsymbol{y}_{mis}$ **do**
7: find distance $\Delta_i = |\hat{\boldsymbol{y}}_{obs,i} - \hat{\boldsymbol{y}}_{mis,k}|$ where $i \neq k$.
8: **randomly sample one donor** from Δ_i with $i = 1, \cdots, 5$ smallest elements and take the corresponding $\hat{\boldsymbol{y}}_{obs}$ to imput $\boldsymbol{y}_{mis}$.

Figure 2 shows how PMM imputes the missing values $\boldsymbol{y}_{imp}$ by randomly select one out of five plausible donors $\boldsymbol{y}_{obs}$ with smallest distance Δ. The $\boldsymbol{y}_{imp}$ has the smallest Δ in this example. PMM has the advantage of imputing real values observed from the data (Schenker and Taylor, 1996, White et al., 2011, Allison, 2015). PMM also gives more robust estimates in the presence of misspecification in the imputation model (Koller-Meinfelder, 2009).

Figure 2: Imputation of a nearest neighbor with PMM



Note. ∘ indicates observed values y_{obs}, • indicates imputed value y_{imp} and ⋄ indicates fitted values $\hat{y}_{obs}$ and $\hat{y}_{mis}$.
Source: adapted from (Koller-Meinfelder, 2009, p.32)

King et al. (2001) propose EMB which combines Expectation Maximisation (EM) algorithm with bootstrap sampling. Unlike PMM, EMB uses predicted values of a linear regression fitted to the observed data to impute missing values. EMB assumes variables in $\boldsymbol{\mathcal{D}}$ are multivariate normal and data are missing at random (King et al., 2001). The imputation formula is

$$\widetilde{\mathcal{D}}^{(j)}_{mis,i} = \boldsymbol{\mathcal{D}}^{(-j)}_{obs,i}\widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\epsilon}}_i, \tag{3}$$

where $\tilde{}$ indicates a random draw from the appropriate posterior. The symbol $\widetilde{\mathcal{D}}^{(j)}_{mis,i}$ denotes a imputed value for row i and column j and $\widetilde{\boldsymbol{\mathcal{D}}}^{(j)}_{obs,i}$ denotes the vector of values observed of all columns in row i except column j. The coefficients $\widehat{\boldsymbol{\beta}}$ from a can be calculated from the complete data parameters be $\vartheta = (\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. The randomness of $\widetilde{\mathcal{D}}^{(j)}_{mis,i}$ is created by both estimation uncertainty due to unknown ϑ and uncertainty in $\widetilde{\boldsymbol{\epsilon}}_i$ because Σ is not a matrix of zero (Honaker and King, 2010).

The box 1 simplifies the notation by removing the superscripts and subscripts for $\boldsymbol{\mathcal{D}}_{mis}$ and $\boldsymbol{\mathcal{D}}_{obs}$ to describe the concept of the algorithm (Tan et al., 2009).

Algorithm 2 EMB algorithm

1: **procedure**

2: generate m bootstrap sample of size n with replacement from the posterior $Pr(\vartheta) \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis}$ described in (9b).

3: keep draws of $\tilde{\vartheta}$ with probabilities proportional to the importance ratio - the ratio of the posterior to the asymptotic normal approximation evalued at $\tilde{\vartheta}$. King et al. (2001) defines the importance ratio (IR) without prior as

$$IR = \frac{\ell(\tilde{\vartheta} \mid \boldsymbol{\mathcal{D}}_{obs})}{\mathcal{N}(\tilde{\vartheta} \mid \tilde{\vartheta}, V(\tilde{\vartheta}))}.$$

4: draw $\widetilde{\boldsymbol{\beta}}$ from a multivariate normal distribution centered at $\widehat{\boldsymbol{\beta}}$ with covariance matrix $\tilde{\sigma}^2(\boldsymbol{\mathcal{X}}_{obs}^{\top}\boldsymbol{\mathcal{X}}_{obs})^{-1}$.

5: in each sample m, fill in $\boldsymbol{\mathcal{D}}_{mis}$ by running an EM algorithm described below.

6: Let $\tilde{\vartheta}^{(i)}$ be the current guess of $\tilde{\vartheta}$

7: Expectation step computes the Q function defined by

$$\begin{aligned} Q(\tilde{\vartheta}^{(i)} \mid \tilde{\vartheta}) &= E[\ell(\tilde{\vartheta}; \boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{D}}_{mis}) \mid \boldsymbol{\mathcal{D}}_{obs}\tilde{\vartheta}^{(i)}] \\ &= \int \ell(\tilde{\vartheta}; \boldsymbol{\mathcal{D}}_{mis}, \boldsymbol{\mathcal{D}}_{obs}) \times f(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}, \tilde{\vartheta}^{(i)}) d\boldsymbol{\mathcal{D}}_{mis}, \end{aligned}$$

8: Maximisation step maximises Q with respect to $\tilde{\vartheta}$ to obtain

$$\tilde{\vartheta}^{(i+1)} = argmax\, Q(\tilde{\vartheta}^{(i)} \mid \tilde{\vartheta}).$$

9: **repeat**both Expectation and Maximisation steps

10: **until** convergence occurs

Baraldi and Enders (2010) discussed how multiple imputation methods create many copies of datasets with different imputed values. These datasets are analysed using the same estimation step to generate multiple sets of parameters and normal standard errors. The final result is derived by using model averaging to incorporate the uncertainty associated with the model selection process into standard errors and confidence intervals (Schomaker and Heumann, 2014). It is unclear if model averaging from multiple imputed datasets provides the best results. This study applies each method 20 times to the 1% sample and we select the best imputed dataset which maximises the likelihood for (1) from the 40 datasets (Fay, 1992, Meng, 1994).

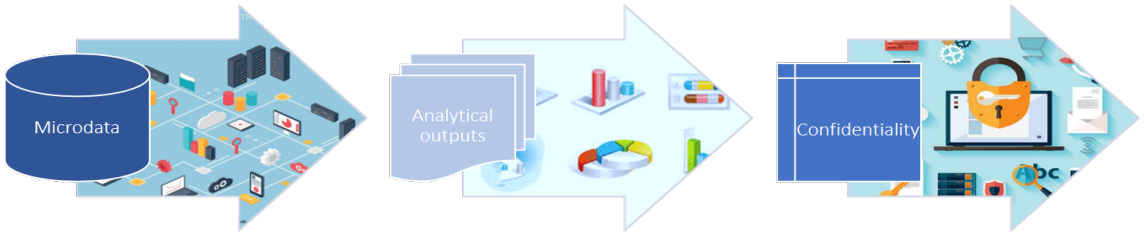
3 DISCLOSURE CONTROL

O'Keefe and Shlomo (2012) categorise statistical disclosure control methodologies into two main approaches - confidentialised input or confidentialised output. Examples of confidentialised input methods include aggregation, geographical suppression, rounding, swapping and adding noise (see Figure 3). However, it is often difficult to quantify the amount of information loss or level of protection achieved using confidentialised input approaches. Rubin (1993) proposed a method to generate synthetic data by repeatedly sampling from a statistical model estimated from actual microdata. The synthetic datasets can be used for inference while protecting confidentiality.

Figure 3: Confidentialise input approach

Confidentialised output approaches allow data access in a remote analysis system. The system takes a query and returns the results to the analyst. The analyst does not have direct access to the microdata. The remote system imposes restrictions on the queries and applies routines to deliver confidentialised results (see Figure 4).

Figure 4: Confidentialise output approach



The study compares confidentialised input and output approaches using imputed microdata. The aim is to compare the protection of confidentiality using different approaches. We explore both synthetic data and perturbation. Reiter (2009) discussed the fully synthetic or partially synthetic data approaches. Consider the following example, an analyst wants to estimate a Cobb-Douglas production function from one of the ABS business surveys. The survey contains 6,500 businesses. Fully synthetic randomly simulates values for business turnover and capital investment for 500 businesses from the joint distributions of the model. These distributions are estimated using the survey data or other relevant information. The result is one synthetic dataset. This process is repeated, each time using a different 500 businesses, to generate multiple synthetic datasets.

In comparison, the partially synthetic approach only replaces sensitive values in the microdata with multiple imputations. Consider the same example and assume the capital investment variable is considered to be sensitive. The statistical agency wants to suppress any income value exceeding $20 million. So any businesses in the sample with investment exceeding the threshold will be excluded in the simulated datasets. The disclosure protection of PS depends on the nature of the synthesis. Replacing identifying variables with imputations makes it very unlikely for users to identify the original values, however it does not guarantee 100% protection. Synthetic data preserves the underlying statistical relationships found in the observed data.

3.1 Synthetic data

This paper explores two synthetic data generation methods for Australian business microdata - the sequential regression (SR) of Raghunathan et al. (2001) and non-parametric imputation based on classification and regression trees (CART) proposed by Reiter (2005b). We use a different font, i.e. $\mathbf{X}$, to represent imputed $N \times 15$ matrix containing all the independent variables in (1). Similarly, we use $\mathbf{y}$ to represent the imputed vector containing dependent variable in (1).

We create fully synthetic data for three variables $\ln y$, $\ln K$ and $\ln M$ from an imputed experimental dataset (see APPENDIX A). The firm output $\ln y$ is the logarithm of total sales adjusted for the repurchase of stocks divided by the total number of employees. Firm capital $\ln K$ is the sum of equipment depreciation, business rental expenses and capital investment deductions divided total number of employees. Material costs $\ln M$ are the inputs used in the production process divided by the total number of employees. These variables have higher disclosure risks because the business information is more sensitive. The synthetic variables combined with the original variables $\ln Firm\_Age$ and time indicator variables to estimate (1).

The SR formula for generating synthetic data for $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \tag{4}$$

where $\mathbf{X}$ is the matrix whose columns contain the observed variables used to predict $\mathbf{y}$ and $\boldsymbol{\beta}$ is the vector of weights given each of the observed variables used to predict $\mathbf{y}$. We apply (4) three times with $\mathbf{y}$ denoting each of the three variables $\ln y$, $\ln K$ and $\ln M$. We use $\mathbf{X}$, $\mathbf{X}^{(K)}$ and $\mathbf{X}^{(M)}$ to denote the matrix for creating synthetic data in $\ln y$, $\ln K$ and $\ln M$, respectively. So if the synthetic data variable is $\ln y$ then $\mathbf{X}$ includes all the independent variables in (1) in APPENDIX A. In comparison, if the synthetic data variable is $\ln K$ then $\mathbf{X}^{(K)}$ includes all the independent variables and $\ln y$ but excludes $\ln K$. Similarly, if the synthetic data variable is $\ln M$ then $\mathbf{X}^{(M)}$ includes all the independent variables and $\ln y$ but excludes $\ln M$.

The SR method uses appropriate regression models for different variable types. For example, continuous variables are generated using a normal model and binary variables using a logit model. This study only creates synthetic data for continuous variables. The SR method generates a continuous vector $\mathbf{y}^{seq}$ from the parameters directly estimated from the fitted regression as follows. First draw a new value $\theta = (\sigma^2, \boldsymbol{\beta})$ from $Pr(\theta \mid \mathbf{y})$. Specifically, the variance is drawn from $\sigma^2 \mid \mathbf{X} \sim (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\chi^{-2}_{n-k}$, where n is the total number of observations and k is the dimension of $\boldsymbol{\beta}$. The coefficients are drawn from $\boldsymbol{\beta} \mid \sigma^2, \mathbf{X} \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$. Second, the synthetic values for $\mathbf{y}^{seq}$ are drawn from the regression model $\mathbf{y}^{seq} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. The imputations are generated for each variable sequentially (Drechsler, 2011).

The CART algorithm estimates the conditional distribution of a univariate outcome given multivariate predictors by partitioning the predictors into groups with similar outcomes. The partitions are created by recursive binary splits of the predictors in a tree structure with leaves. The values in each leaf represent the conditional distribution of outcomes that satisfy the partitioning criterion. Effectively, CART preserves the underlying relationships between variables by creating models with many interaction effects (Reiter, 2005b, Burgette and Reiter, 2010).

To create $\mathbf{y}^{cart}$, we first fit a tree relating $\mathbf{y}$ to $\mathbf{X}$. We do this separately for all three variables $\ln y$, $\ln K$ and $\ln M$. The algorithm minimises the deviation of $\mathbf{y}$ within each leaf and stops splitting when the deviation is below 0.001. We do this for three variables and label these trees $tree^{(y),(K),(M)}$. We use $\mathbf{y}_{leaf}$ to represent the predicted values of terminal leaves $leaf^{(y),(K),(M)}$ in the trees. In each leaf of the tree, we use the Bayesian bootstrap to draw new values from $\mathbf{y}_{leaf}$ to create synthetic data (Reiter, 2005b). The Bayesian bootstrap differs from the standard bootstrap by varying the selection probabilities in the re-sampling process (Rubin, 1981). The main advantage of using the Bayesian bootstrap is adding uncertainty in each leaf because the number of values in each leaf tend to be small (Reiter, 2005b).

We generate 20 **synthetic datasets using each method. We use these datasets to fit** (1) **and choose the synthetic dataset with highest log-likelihood for the analysis.**

3.2 Perturbation

The *confidentialised input approach* produces synthetic microdata that allows researchers to analyse the microdata. In comparison, the *confidentialised output approach*, e.g. perturbation, does not allow researchers to access the underlying microdata. Researchers can only explore data and perform modelling analyses within a secured remote environment. In this environment, on-the-fly routines are applied to confidentialise results for analysis. These routines protect confidentiality while maximising the utility of the microdata.

The perturbation algorithm starts by considering the estimation for model (1) as solving $Sc(\boldsymbol{\alpha}; \mathbf{X}; \mathbf{y}) = 0$, where $Sc(\boldsymbol{\alpha}; \mathbf{X}; \mathbf{y}) = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\alpha})$. The algorithm then adds the noise $\mathbf{e}$ to the score function. We use $\boldsymbol{\alpha}^{pert}$ to denote the coefficients after the score function has been perturbed. The perturbed estimating equation can be expressed as

$$Sc(\boldsymbol{\alpha}^{pert}; \mathbf{X}; \mathbf{y}) = \mathbf{e}. \tag{5}$$

The amount of perturbation is based on a record's contribution to the coefficients in the estimating equation. The perturbation is added using $\mathbf{e} = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\alpha})\mathbf{u}$, where noise $\mathbf{u}$ is generated independently from the uniform distribution with the range $(-1, 1)$. The estimated coefficients after perturbation are $\widehat{\boldsymbol{\alpha}}^{pert} = \widehat{\boldsymbol{\alpha}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{e}$ where $\widehat{\boldsymbol{\alpha}}$ is the estimated coefficient using the original microdata. The solution of (5) $\widehat{\boldsymbol{\alpha}}^{pert}$ is an unbiased estimate of $\boldsymbol{\alpha}$ because the noise is small and its expected value has mean zero $E(\mathbf{e}) = 0$. The perturbation has a similar effect to removing records that have large contribution to the estimated coefficients (Chipperfield and O'Keefe, 2014, Chipperfield, 2014).

4 EMPIRICAL RESULTS

Analysing business microdata can lead to re-identification because it contains large business units, unlike household or person microdata which have a large number of similar respondents. This means protecting confidentiality of business microdata can be different from protecting household or person microdata. Statistical agencies often take strong protection measures to minimise the likelihood of disclosure (O'Keefe and Shlomo, 2012). There are different approaches to estimate disclosure risks and one approach is to estimate risk scores for individual records using the probability of matching between sample and population microdata. These individual record risk scores are then aggregated for the entire data file, see (Bethlehem et al., 1990, Shlomo, 2010, Drechsler, 2011).

The confidentialised output approach does not generate individual confidentialised units so it is not feasible to calculate the individual risk score. Instead, we follow the approach of O'Keefe and Shlomo (2012) and show the absolute differences between the identifying variables in the confidentialised and original microdata across selected industries in the disclosure risk models. $\ln y$, $\ln K$ and $\ln M$ on the logarithms of the total number of employees in each firm j in these models. There are strong positive correlations between firm size and variables with higher disclosure risks such as $\ln y, \ln K$ and $\ln M$. Figures 5 and 7 in APPENDIX C shows that the synthetic data approach generally provides more confidentiality protection as the absolute differences $|\delta|$ are often wider than in the perturbation approach.

This study compares the estimated coefficients using confidentialised input and outputs approaches with the estimated coefficients using original microdata for measuring utility. Figures 9, 10, 11 and 12 in APPENDIX C compare the estimated coefficients using

different approaches for selected industries. The Figures show the results for the main variables and intercepts. In the large sample size, the sequential regression provides the best protection but with a higher utility trade off. The estimated coefficients using the perturbation approach have relatively smaller biases but the results are comparable. In comparison, the results are mixed in small sample sizes in industries like mining or public administrative. In general, the standard deviations for the coefficients are larger using the synthetic data approach. The coefficient plots for the rest of the variables can be found in the APPENDIX *C*, see Figures 13, 14, 15 and 16.

Figure 17, in the APPENDIX *D*, shows the model residuals using hex-bin plots. Figure 18 shows the quantile-quantile normal plots for all industries. There are no notable differences when we compare different approaches with the model results using unconfidentialised data. However, the analysis shows that there are differences when we consider mining industry. The confidentialised inputs approach provides better protection with trade off in variance estimation see Figure 19 and Figure 20 in APPENDIX *D*.

5 CONCLUSIONS

This research compares synthetic data and perturbation approaches for disseminating Australian business microdata. The preliminary results show that synthetic data can be a possible dissemination tool to make more business microdata accessible while ensuring confidentiality.

The analysis shows that the confidentialised input approach provides more protection than the confidentialised output approach in this particular setting - one percent sample file of business microdata. This is partly because the researchers have access to the microdata so there is a stronger need to add more noise for protection. The amounts of utility loss from synthetic data and perturbation approaches are comparable because the estimated coefficients are similar. Synthetic data could be a possible approach for the ABS to consider to enhance access to business microdata. This preliminary research has several areas for possible extension including:

- exploring multilevel models for creating synthetic data to better capture the hierarchical structure of the dataset (Drechsler, 2015).
- considering other non-parametric approaches for synthetic data such as random forest or differential privacy (Drechsler and Reiter, 2011).
- exploring synthetic data approaches which also maintain differential privacy standard (Sarwate and Chaudhuri, 2013).

6 ACKNOWLEDGEMENTS

Authors would like to express our gratitude to the following ABS colleagues - Dr Siu-Ming Tam, Sybille McKeown, Lisette Aaron, Diane Braskic, Dr Philip Gould, Rowan Hatley, Dr Sarah Hinde, Grace Kim, David Taylor, Liza Tiy and Carter Wong for their helpful comments and support for this research. Dr Jörg Drechsler for sharing R code and the ABS Remote Execution Environment for Microdata project team for the developing the R code and Sebastian Lucie's advice and those who provided comments at the Synthetic Datasets for Statistical Disclosure Control - Research and Applications Around the World workshop at the 61st ISI World Statistics Congress 2017. We remain solely responsible for the views expressed in this paper.

7 REFERENCES

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002), "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Report, US Census Bureau, access at link on 12022016.

Abowd, J. M. and Schmutte, I. M. (2015), "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods," Access at link on 12022017.

ABS (2016), "Information Paper Transforming Statistics for the Future," Access at link on 01022017.

— (2017), "Microdata Entry Page," Access at link on 01082017.

Allison, P. (2015), "Imputation by Predictive Mean Matching Promise and Peril," Access at link on 12042016.

Baraldi, A. N. and Enders, C. K. (2010), "An introduction to modern missing data analyses," *Journal of School Psychology*, 48, 5–37.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," *Journal of the American Statistical Association*, 85, 38–45.

Blackmore, K. and Nesbitt, K. (2013), "Verifying the Miles and Snow strategy types in Australian small- and medium-size enterprises," *Australian Journal of Management*, 38, 171–190.

Breunig, R. and Wong, M.-H. (2008), "A Richer Understanding of Australia's Productivity Performance in the 1990s: Improved Estimates Based Upon Firm-Level Panel Data," *Economic Record*, 84, 157–176.

Breunig, R. V. and Bakhtiari, S. (2013), "Outsourcing and Innovation: An Empirical Exploration of the Dynamic Relationship," *The B.E. Journal of Economic Analysis & Policy*, 13, 395.

Burgette, L. F. and Reiter, J. P. (2010), "Multiple imputation for missing data via sequential regression trees," *American Journal of Epidemiology*, 172, 1070–1076.

Chien, C.-H. and Mayer, A. (2015), "Use of a prototype Linked Employer-Employee Database to describe characteristics of productive firms," Report, Australian Bureau of Statistics, access at link on 05082016.

Chien, C.-H., Welsh, A. H., and Breunig, R. V. (2019), "Approaches to Analysing Micro-Drivers of Aggregate Productivity," Report, Australian Bureau of Statistics, access at online.

Chipperfield, J. O. (2014), "Disclosure-protected inference with linked microdata using a remote analysis server," *Journal of Official Statistics*, 30, 123–146.

Chipperfield, J. O. and O'Keefe, C. M. (2014), "Disclosure-protected Inference Using Generalised Linear Models," *International Statistical Review*, 82, 371–391.

Desai, T., Ritchie, F., and Welpton, R. (2016), "Five Safes: designing data access for research," .

Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control Theory and Implementation*, Lecture Notes in Statistics, New York: Springer.

— (2015), "Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity," *Journal of Educational and Behavioral Statistics*, 40, 69–95.

Drechsler, J. and Reiter, J. P. (2011), "An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets," *Computational Statistics & Data Analysis*, 55, 3232–3243.

Fay, R. E. (1992), "When Are Inferences from Multiple Imputation Valid?" in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 227–232.

Harel, O. and Zhou, X. (2007), "Multiple imputation: review of theory, implementation and software," *Statistics in medicine*, 26, 3057–3077.

Healy, J., Mavromaras, K., and Sloane, P. J. (2015), "Adjusting to skill shortages in Australian SMEs," *Applied Economics*, 47, 2470–2487.

Honaker, J. and King, G. (2010), "What to do about missing values in time-series cross-section data," *American Journal of Political Science*, 54, 561–581.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001), "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, 95, 49–69.

Koller-Meinfelder, F. (2009), "Analysis of incomplete survey data-multiple imputation via bayesian bootstrap predictive mean matching," Thesis.

Little, R. J. (1988), "A test of missing completely at random for multivariate data with missing values," *Journal of the American statistical Association*, 83, 1198–1202.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley and Sons.

Loong, B. (2012), "Topics and applications in synthetic data," Thesis.

Mare, D. C., Hyslop, D. R., and Fabling, R. (2016), "Firm productivity growth and skill," *New Zealand Economic Papers*, 1–25.

Meng, X.-L. (1994), "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, 538–558.

Nguyen, T. and Hansell, D. (2014), "Firm dynamics and productivity growth in Australian manufacturing and business services Oct 2014," Report, ABS, access at online.

O'Keefe, C. M. and Shlomo, N. (2012), "Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data," *Trans. Data Privacy*, 5, 403–432.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, 27, 85–96.

Reiter, J. P. (2005a), "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 185–205.

— (2005b), "Using CART to generate partially synthetic public use microdata," *Journal of Official Statistics*, 21, 441.

— (2009), "Multiple Imputation for Disclosure Limitation Future Research Challenges," *Journal of Privacy and Condentiality*, 1, 223–233.

Rubin, D. B. (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

— (1993), "Statistical disclosure limitation," *Journal of official Statistics*, 9, 461–468.

Sarwate, A. D. and Chaudhuri, K. (2013), "Signal Processing and Machine Learning with Differential Privacy: Algorithms and Challenges for Continuous Data," *IEEE Signal Processing Magazine*, 30, 86–94.

Schenker, N. and Taylor, J. M. G. (1996), "Partially parametric techniques for multiple imputation," *Computational Statistics and Data Analysis*, 22, 425–446.

Schomaker, M. and Heumann, C. (2014), "Model selection and model averaging after multiple imputation," *Computational Statistics and Data Analysis*, 71, 758–770.

Shlomo, N. (2010), "Releasing microdata Disclosure risk estimation, data masking and assessing utility," *Journal of Privacy and Confidentiality*, 2, 7.

Tam, S.-M., Farley-Larmour, K., and Gare, M. (2009), "Supporting research and protecting confidentiality. ABS microdata access: Current strategies and future directions," *Statistical Journal of the IAOS*, 26, 65–74.

Tan, M. T., Tian, G.-L., and Ng, K. W. (2009), *Bayesian missing data problems: EM, data augmentation and noniterative computation*, Chapman and Hall/CRC.

The Australian Government (1983), "STATISTICS DETERMINATION - REG 7," Access at link on 20012019.

Vink, G., Frank, L. E., Pannekoek, J., and Van Buuren, S. (2014), "Predictive mean matching imputation of semicontinuous variables," *Statistica Neerlandica*, 68, 61–90.

White, I. R., Royston, P., and Wood, A. M. (2011), "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, 30, 377–399.

A SUMMARY STATISTICS

Table 1: Summary Statistics - Test Data

Statistic	N	Mean	St. Dev.	Min	Max
$\ln Firm\_Age$	47,160	7	5	1	20
$\ln \hat{z}_t^{(jk)}$	47,160	10	1	3	15
$\ln y_{jkt}$	47,160	11	1	1	18
$\ln K_{jkt}$	47,160	8	1	−3	16
$\ln M_{jkt}$	47,160	10	2	−3	18
ABN	47,160				
year	47,160	2008	3	2002	2013

[1] $\ln Firm\_Age$ is the logarithm of firm age. Firm age is derived as the current year minus the year of incorporation.
[2] $\ln \hat{z}_t^{(jk)}$ the logarithm of labour inputs.
[3] $\ln y_{jkt}$ is logarithm of per employee value added (i.e. sales adjusted for repurchase of stock) deflated by industry Gross Value Added implicit price deflators
[4] $\ln K_{jkt}$ is the logarithm of per employee cost of capital that includes depreciation, capital rental expenses and capital work deductions deflated by the industry consumption of fixed capital implicit price deflators.
[5] $\ln M_{jkt}$ is logarithm of per employee material costs deflated by Producer Price Indexes Intermediate Goods.

B A BAYESIAN FRAMEWORK FOR IMPUTATION

We assume data are missing at random. The consequence of this assumption is that missing data can be imputed from fitting model on the observed data. The complete data parameters are $\vartheta = (\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. The likelihood of these parameters given the observed data can be expressed as

$$Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) = \int Pr(\boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{R}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \tag{6a}$$

$$= \int Pr(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{R}}, \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis}. \tag{6b}$$

Using Bayes' theorem we can rewrite the first term $Pr(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{R}}, \vartheta)$ in (6b) as $Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}, \vartheta) / Pr(\boldsymbol{\mathcal{R}} \mid \vartheta)$. Substituting the new term into (6b) we have

$$Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) = \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}, \vartheta) d\boldsymbol{\mathcal{D}}_{mis}. \tag{7}$$

Assuming the data are missing at random, the patterns of missing data depend only on the observed data, so (7) is simplified to

$$\begin{aligned} Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) &= \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}, \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \\ &= \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}) \\ &= Pr(\boldsymbol{\mathcal{D}}_{obs} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}). \end{aligned} \tag{8}$$

Maximising (6a) over ϑ is the same as maximising the first term in (8) over ϑ. The likelihood can therefore be expressed as $L(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) \propto Pr(\boldsymbol{\mathcal{D}}_{obs} \mid \vartheta)$. Harel and Zhou (2007) describe the posterior distribution to draw imputations is

$$Pr(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}) = \int Pr(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}, \vartheta) Pr(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) d\vartheta, \text{ where} \tag{9a}$$

$$Pr(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) \propto Pr(\vartheta) \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \tag{9b}$$

is the observed posterior distribution for ϑ and $Pr(\vartheta)$ is an uninformative Jeffreys's prior for Σ.

Figure 5: Disclosure measures - All industries

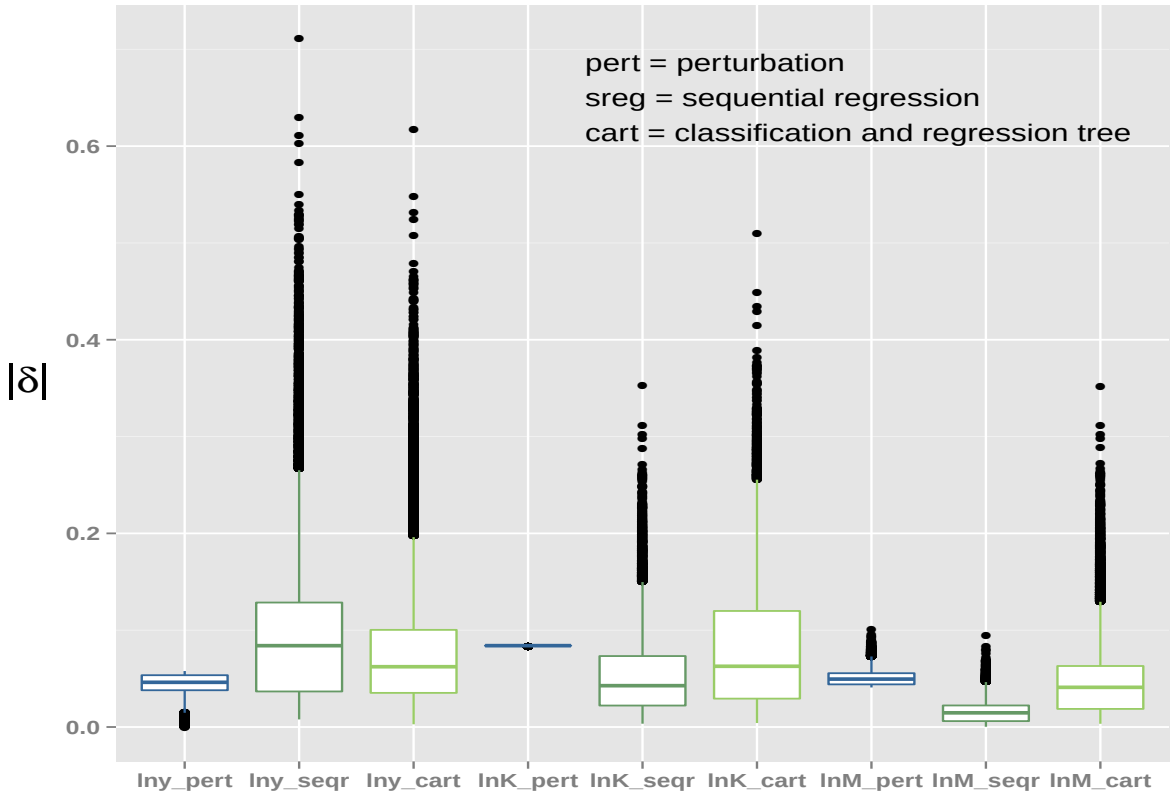


Figure 6: Disclosure measures - Mining

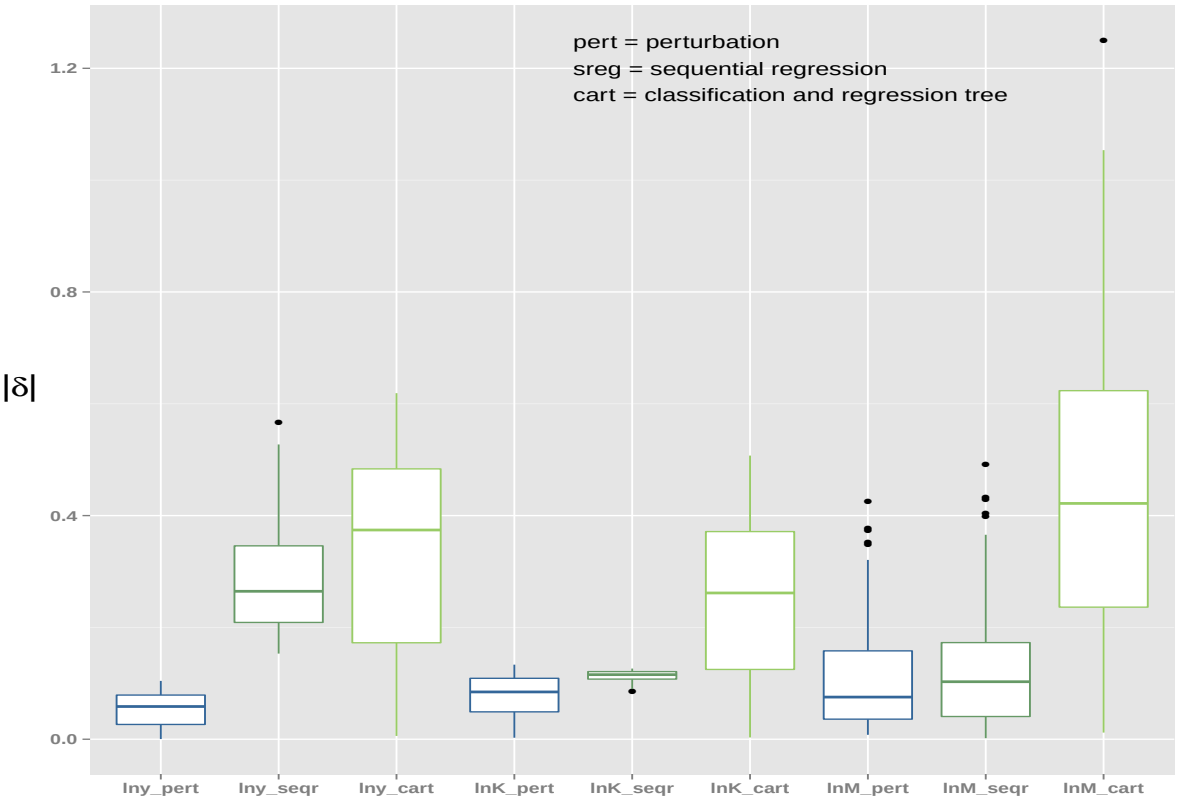


Figure 7: Disclosure measures - Construction

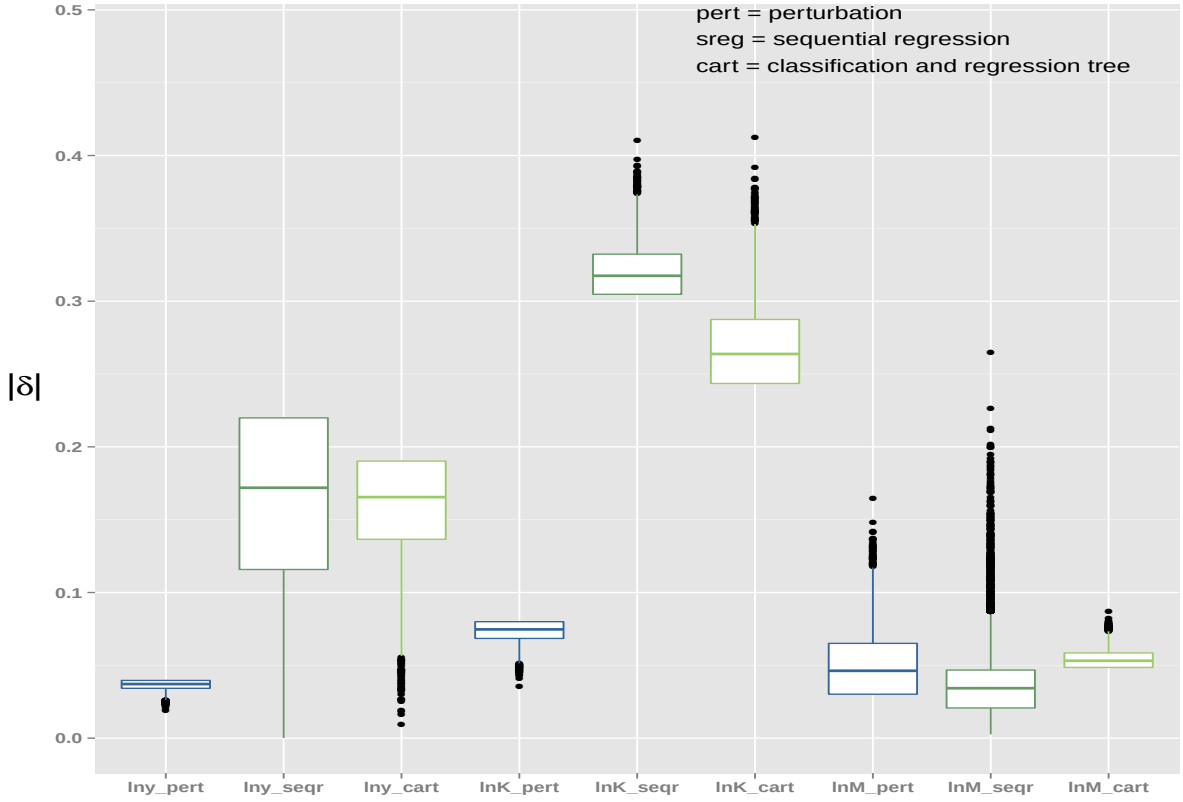


Figure 8: Disclosure measures - Public administration

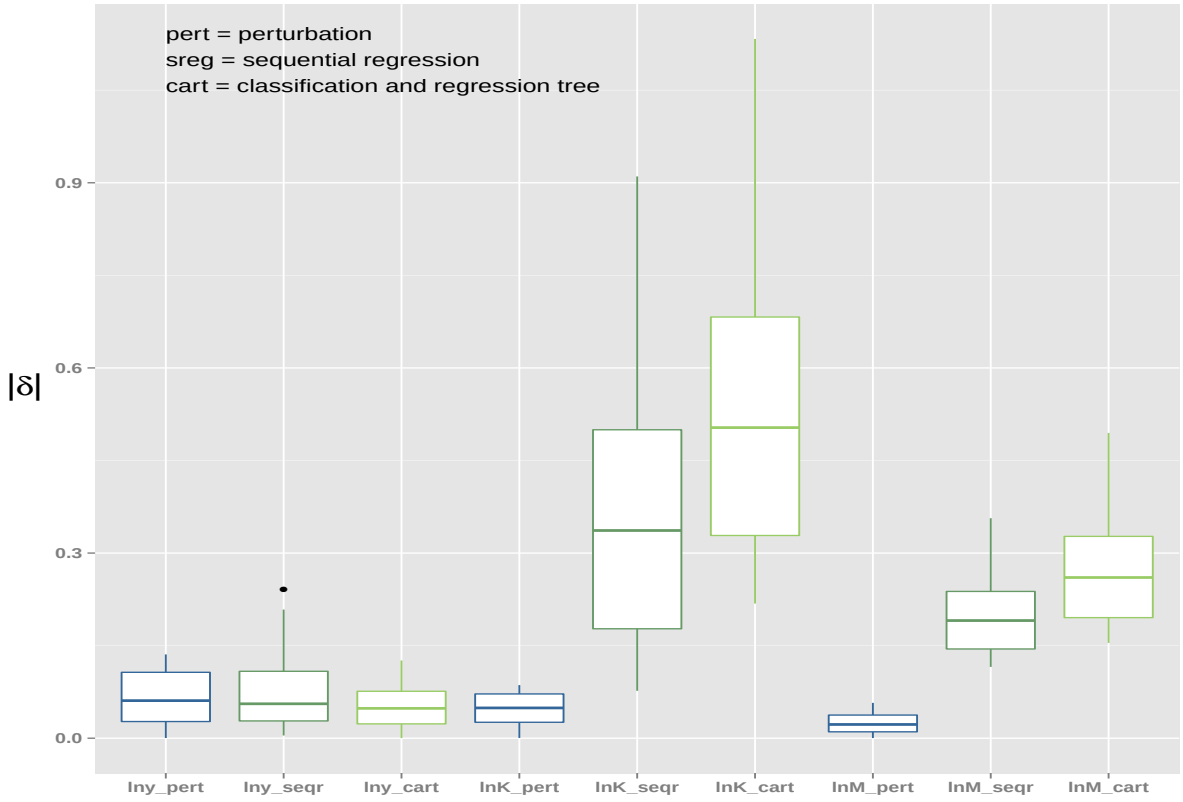


Figure 9: Coefficients plots - All industries

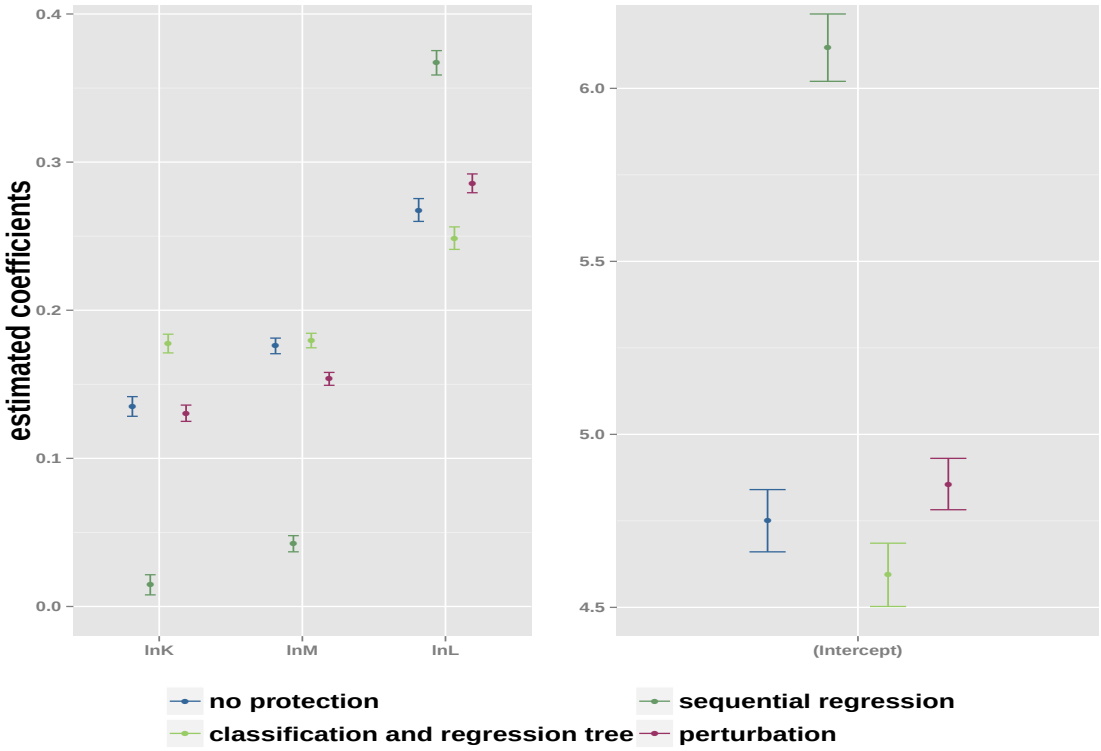


Figure 10: Coefficients plots - Mining

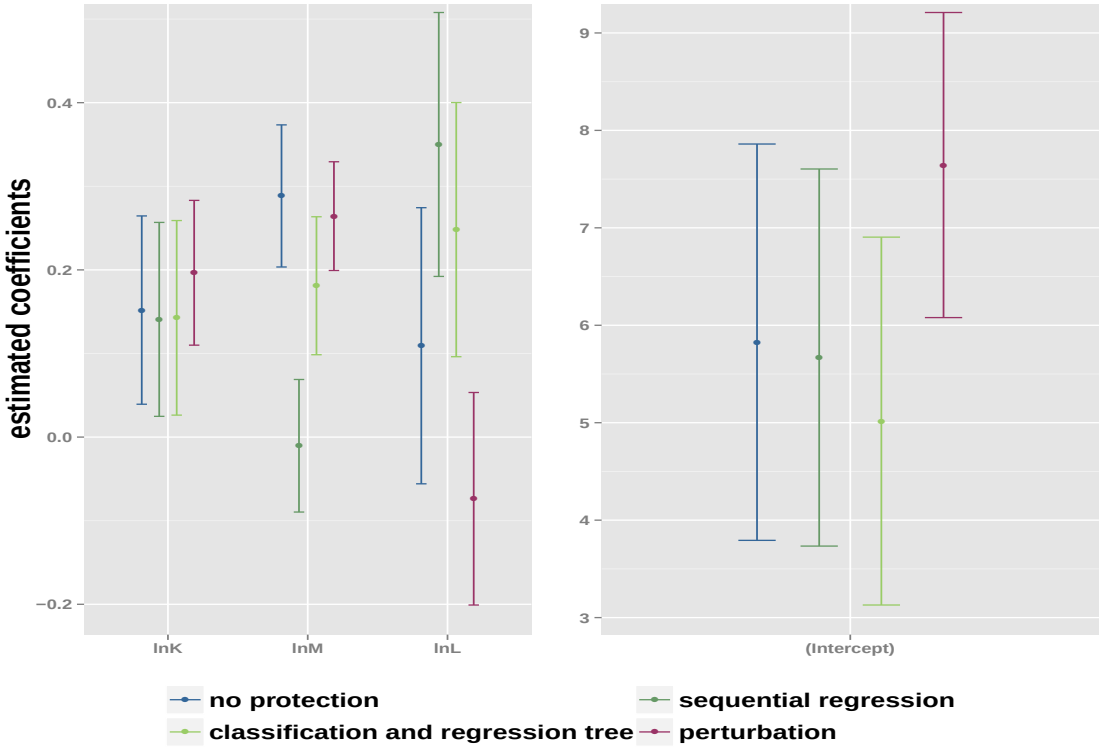


Figure 11: Coefficients plots - Construction

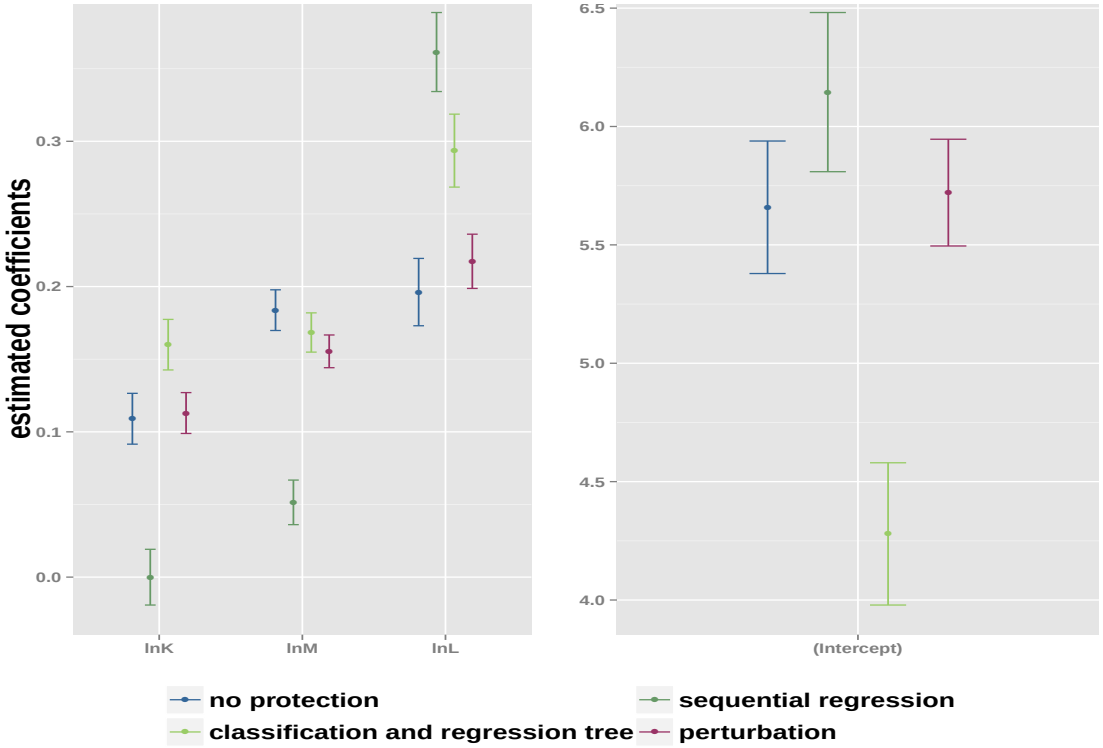


Figure 12: Coefficients plots - Public administration

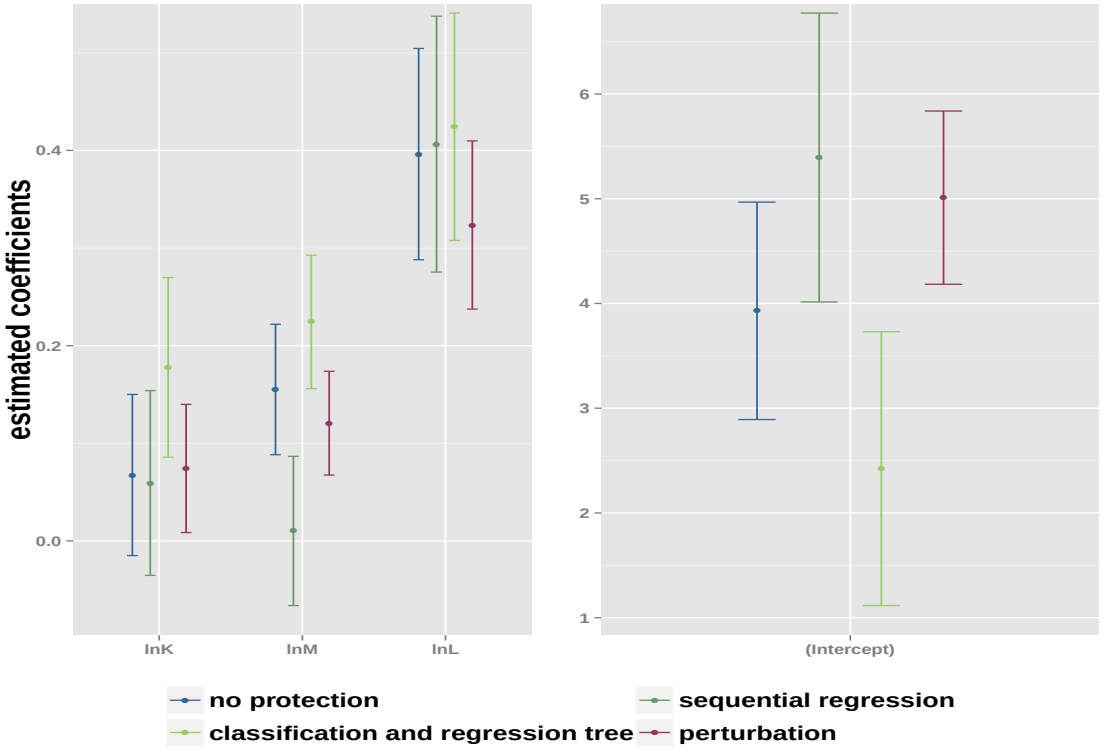


Figure 13: Coefficients plots - All industries

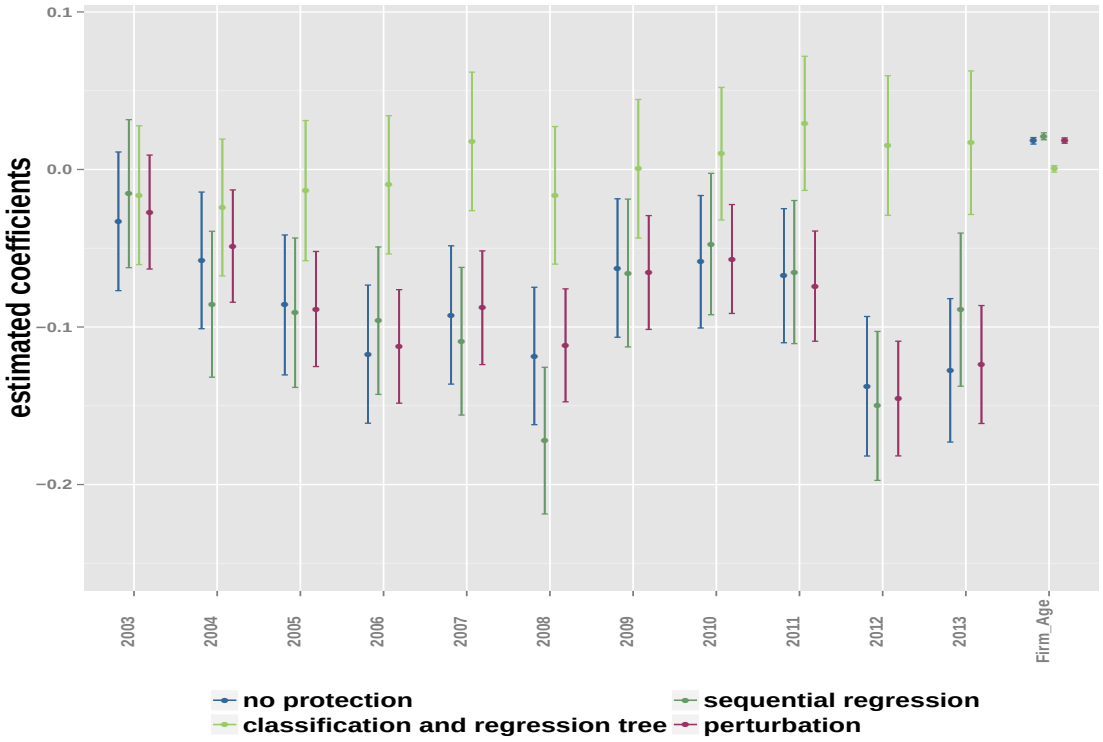


Figure 14: Coefficients plots - Mining

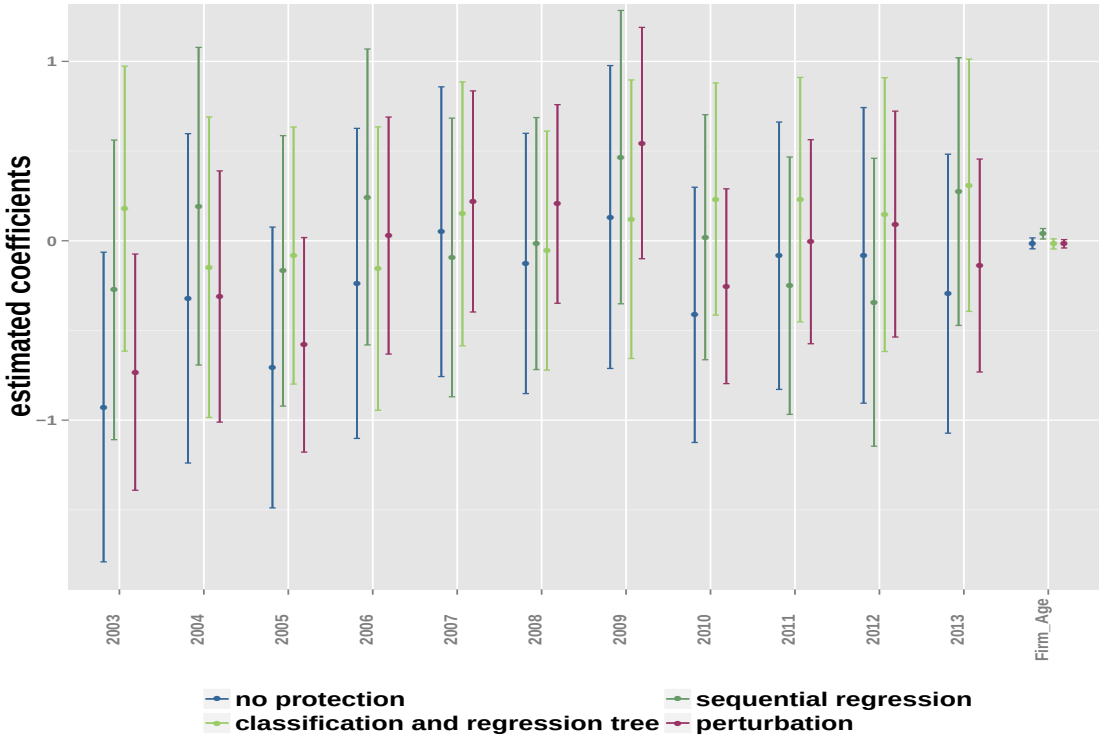


Figure 15: Coefficients plots - Construction

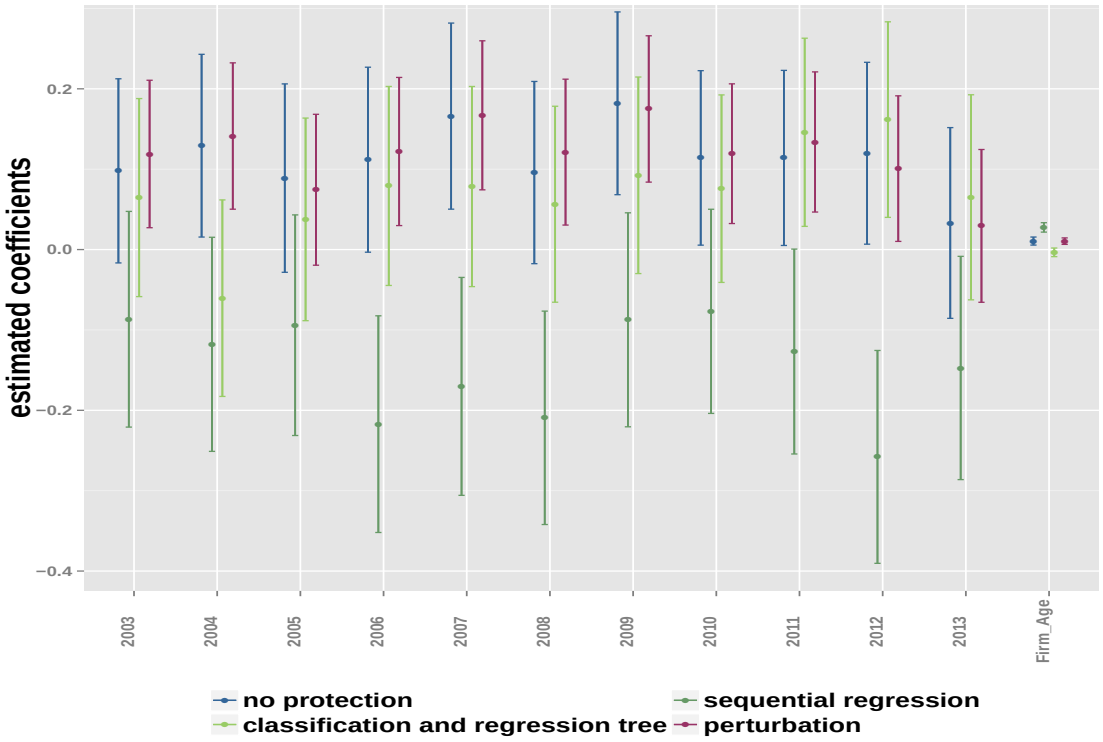


Figure 16: Coefficients plots - Public administration

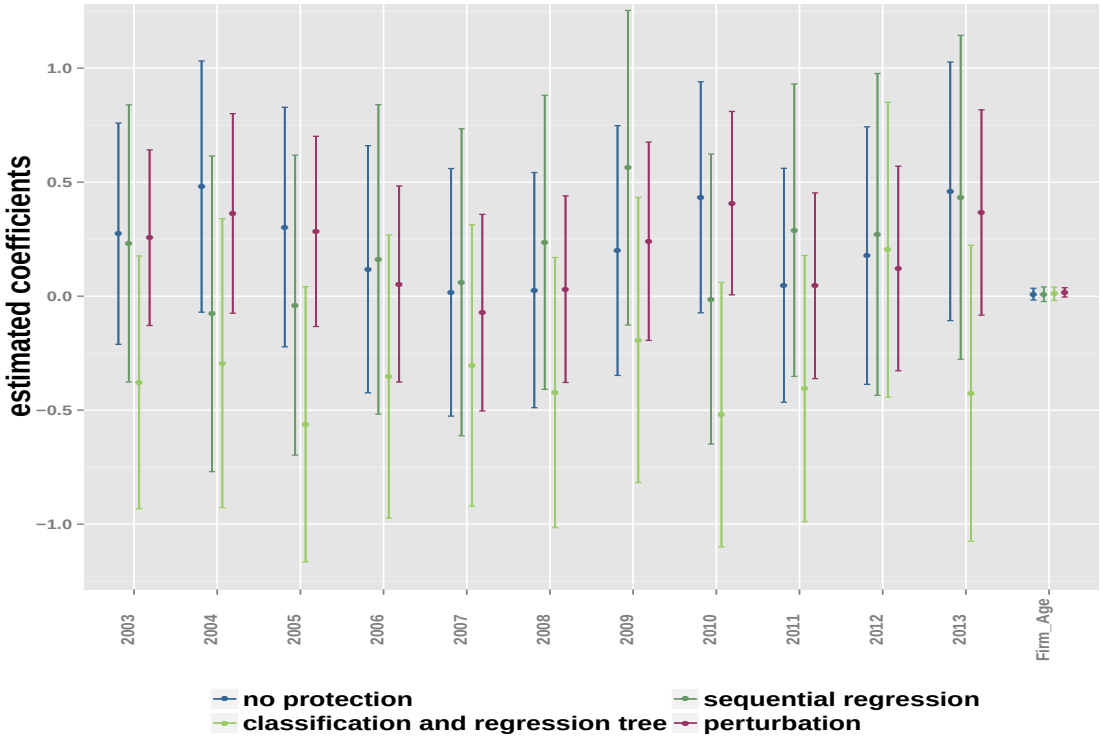
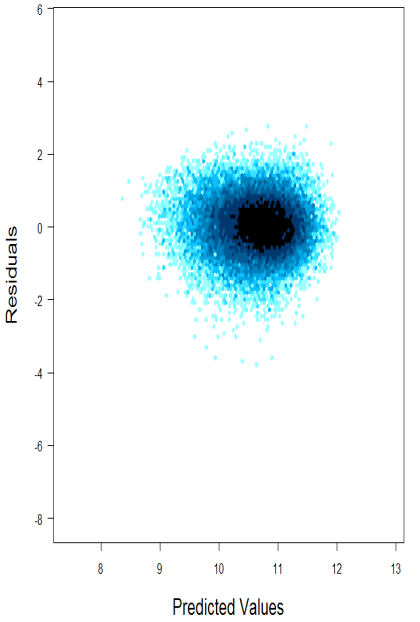
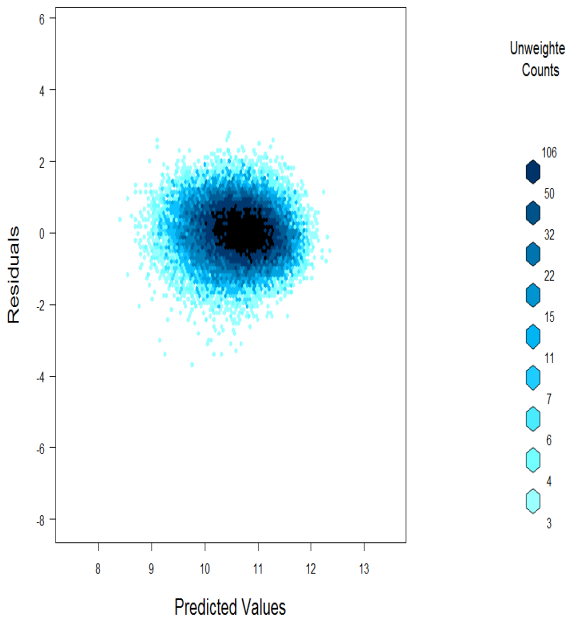


Figure 17: Confidentialised residual plots - ALL industries

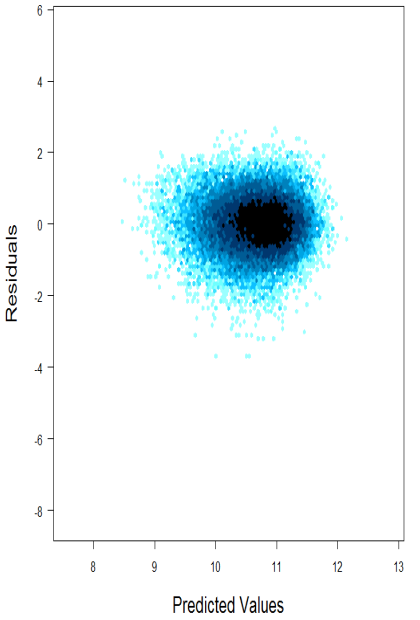
(1) original



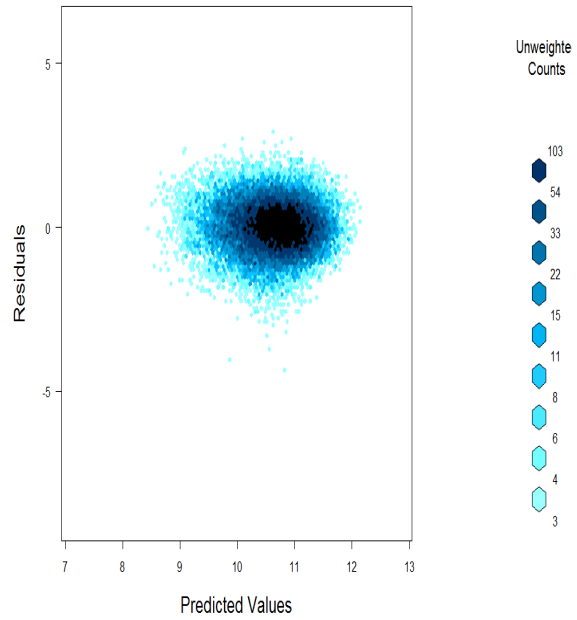
(3) synthetic data - SR



(2) perturbation



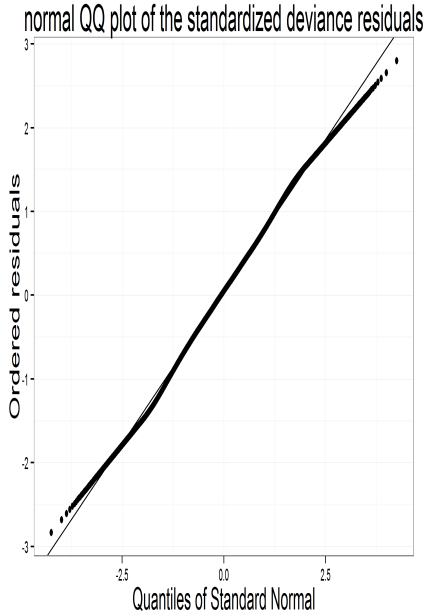
(4) synthetic data - CART



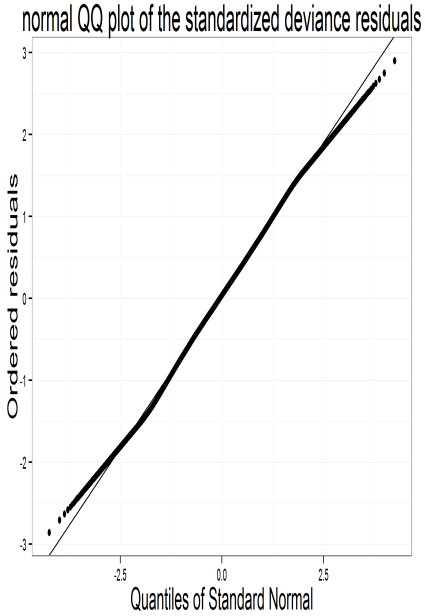
Note. Residuals come from fitting (1) to different approaches. The plotting region on these figures is broken into a mesh of tessellating hexagons, each of which is coloured indicating how many observations lie in that hexagon.

Figure 18: QQ Norm plots - ALL industries

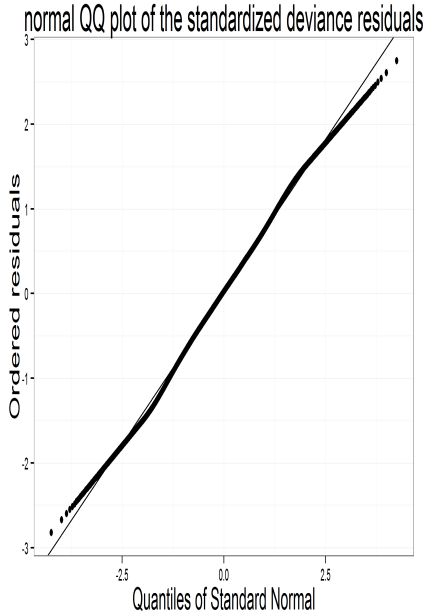
(1) original



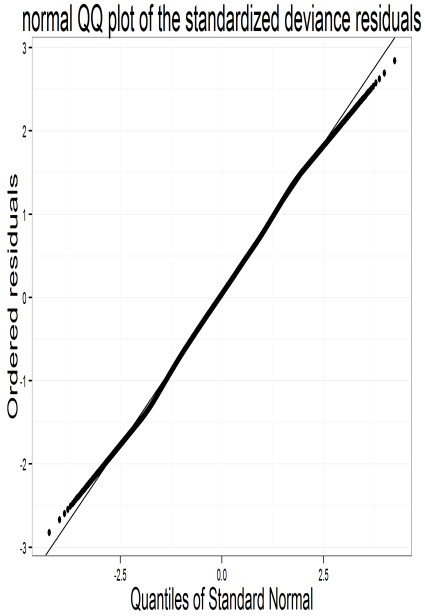
(3) synthetic data - SR



(2) perturbation



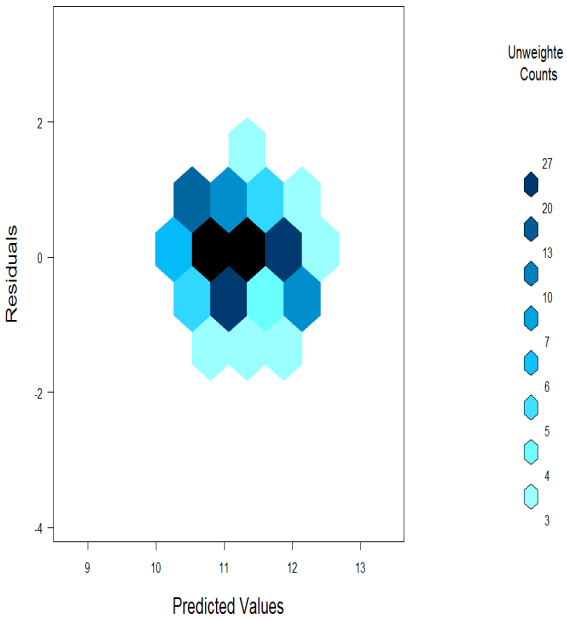
(4) synthetic data - CART



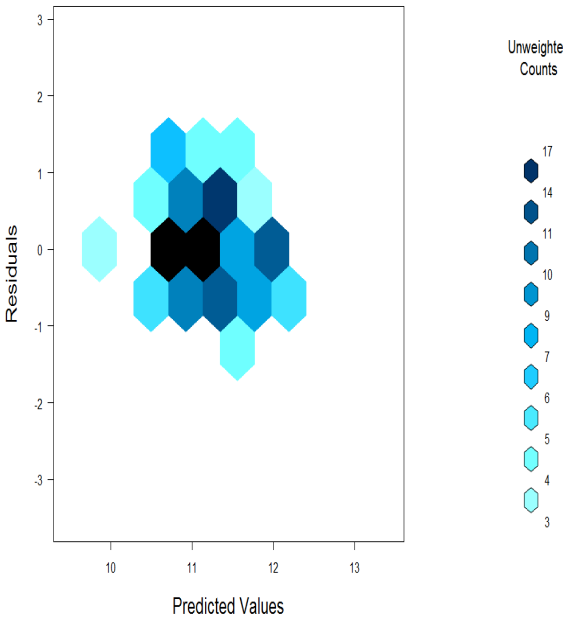
Note. Residuals come from fitting (1) to different approaches. A 45 degree line indicates that residuals are normally distributed.

Figure 19: Confidentialised residual plots - mining

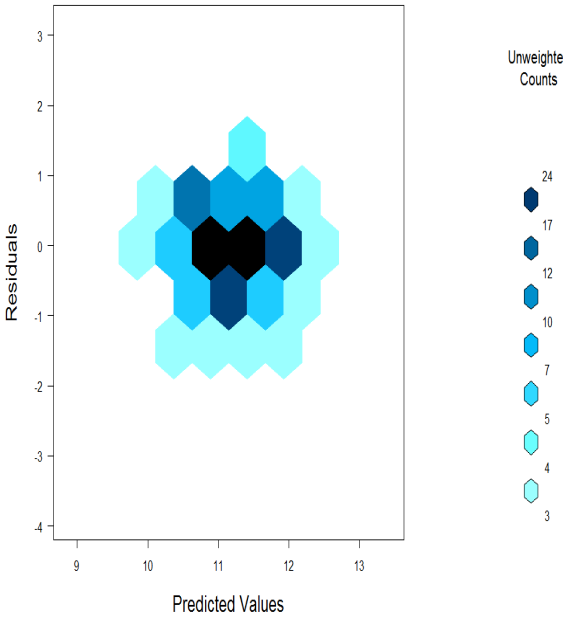
(1) original



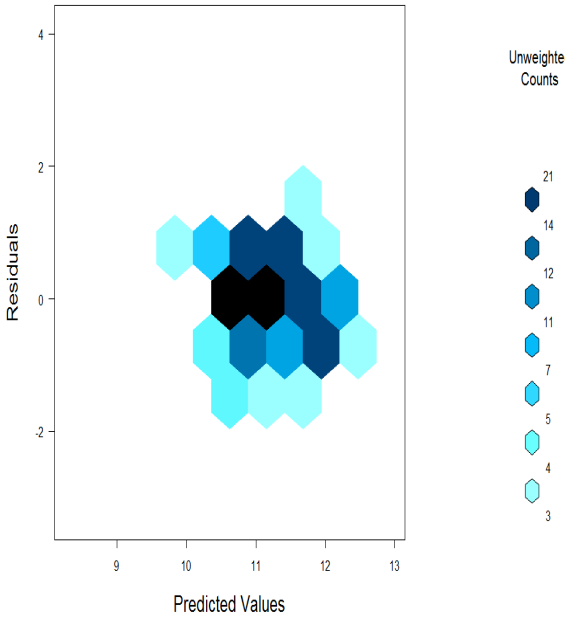
(3) synthetic data - SR



(2) perturbation



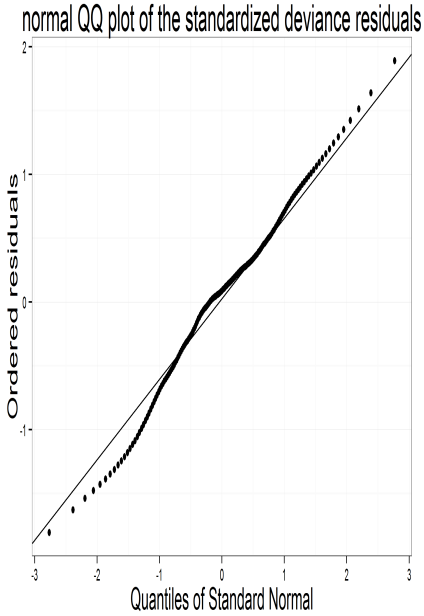
(4) synthetic data - CART



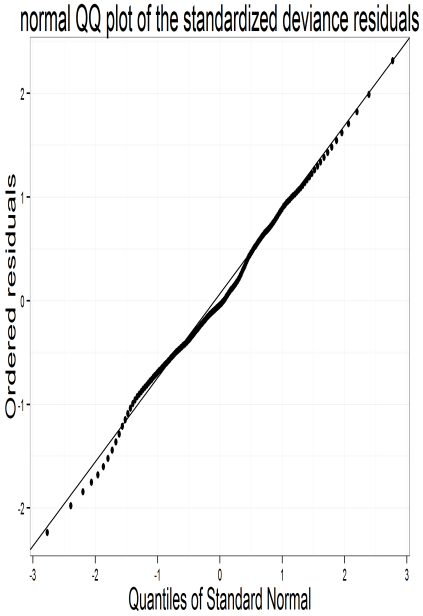
Note. Residuals come from fitting (1) to different approaches. The plotting region on these figures is broken into a mesh of tessellating hexagons, each of which is coloured indicating how many observations lie in that hexagon.

Figure 20: QQ Norm plots - mining

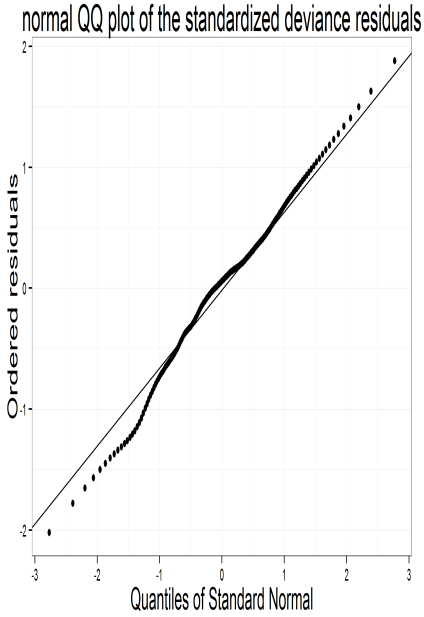
(1) original



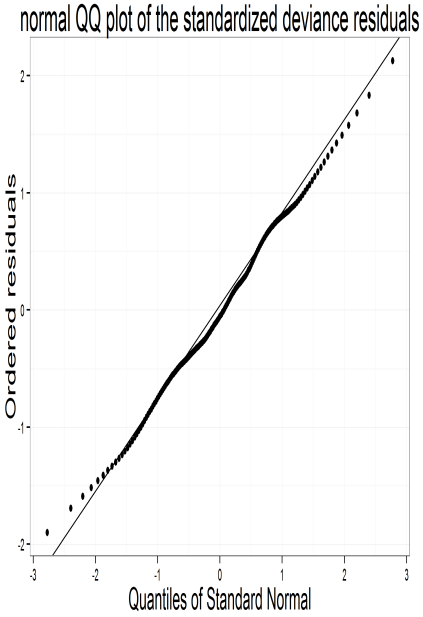
(3) synthetic data - SR



(2) perturbation



(4) synthetic data - CART



Note. Residuals come from fitting (1) to different approaches. A 45 degree line indicates that residuals are normally distributed.

FOR MORE INFORMATION . . .

www.abs.gov.au the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics